Anticipatory Measure for Auction Fraud Detection in Online

Narasamma S, Suma Latha. K, Suma Latha. M

***Abstract*— This paper introduces and presents the Online Modeling of Proactive Moderation System for Auction Fraud Detection by Using online feature selection, stochastic search variable selection (SSVS),coefficient bounds from human knowledge and multiple instance learning. An important usability goal of proactive moderation systems is by applying expert knowledge, such as bounding the rule based feature weights to be positive and multiple instance learning, can significantly improve the performance in terms of detecting more frauds and reducing customer complaints given the same workload from human experts.**

***Keywords*— Online Auction, Fraud Detection, Online Modeling, Online Feature Selection, Multiple Instance Learning.**

I. INTRODUCTION

To start with we focus on the traditional online shopping business model allows sellers to sell a product or service at a preset price, where buyers can choose to purchase if they find it to be a good deal. Online auction however is a different business model by which items are sold through price bidding. There is often a starting price and expiration time specified by the sellers. Once the auction starts, potential buyers bid against each other, and the winner gets the item with their highest winning bid.

We propose an online probit model framework which takes online feature selection, coefficient bounds from human knowledge and multiple instance learning into account simultaneously. By empirical experiments on a real-world online auction fraud detection data we show that this model can potentially detect more frauds and significantly reduce customer complaints compared to several baseline models and the human-tuned rule-based system.

Human experts with years of experience created many rules to detect whether a user is fraud or not. If the fraud score is above a certain threshold, the case will enter a queue for further investigation by human experts. Once it is reviewed, the final result will be labeled as Boolean, i.e. fraud or clean. Cases with higher scores have higher priorities in the queue to be reviewed. The cases whose fraud score is below the threshold value is determined as clean by the system without any human judgment.

In this paper we study the problem of building online models for the auction fraud detection moderation system, which essentially evolves dynamically over time. We propose a Bayesian online model framework for the binary response.

We apply the stochastic search variable selection (SSVC), a well known technique in stastical literature, to handle the dynamic evolution of the feature importance in a principled Way. Note that we are not aware of any previous work that tries to embed SSVS into online modeling.

We consider the expert knowledge to bind the rule-based coefficients to be positive. Finally, we consider to combine this online model with multiple instance learning that gives even better empirical performance. We report the performance of all the above models through extensive experiments using fraud detection datasets from a major online auction website in Asia.

II. LITERATURE SURVEY

C. Chua and J. Wareham. Created the auction fraud taxonomy. According to Chua and Wareham, all the types of fraud listed are very damaging to Internet auction houses. They undermine user trust, which is disastrous for business. D.Gregg and J. Scott discovered that Internet auction fraud takes various forms, such as delivering goods that are different, of low quality, without ancillary components, defective, damaged or black market items. Y. Ku, Y. Chen, and C. Chiu.

Note that while both buyers and sellers can be victims of fraud, a buyer is more easily targeted than a seller. They observed that 89% of seller-buyer pairs conducted just one transaction during the time period of their study; at most, there were four transactions between a seller-buyer pair. This means that the repeated transaction rate for the same seller-buyer pair is lower than 2%. If the transaction rate is much higher than 2%, then the transactions between the seller-buyer pair are suspect and could involve bid shilling or bid shielding.

II. METHODOLOGY

There is a vast range of business models for online consumer auctions. There are over 200 auction sites on the Web, ranging from the free-standing eBay, which handles 87% of online auction transactions, to the auction sites attached to portals like Yahoo! and MSN, to the auction sites attached to e-commerce sites like Amazon, to specialty auction sites. Some sites charge to list items, others do not (although Yahoo! recently started charging for listings). This variety of business models results in a wide range of practices, which are described in detail below. But despite this variety, some general observations can be made about online auctions.

Manuscript published on 30 August 2013.
***** Correspondence Author (s)

Narasamma. S*, Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India.
Sumalatha. K, Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India.
Sumalatha. M, Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India.

Our application is to detect online auction frauds for a major Asian site where hundreds of thousands of cases posted every day. Every new case is sent to the anticipatory moderation system for pre-screening to assess the risk of being fraud. The current system is featured by:

A. RULE-BASED FEATURES:

Human experts with years of experience created many rules to detect whether a user is fraud or not. An example of such rules is “blacklist”, i.e. whether the user has been detected or complained as fraud before. Each rule can be regarded as a binary feature that indicates the fraud likeliness.

B. LINEAR SCORING FUNCTION:

The existing system only supports linear models. Given a set of coefficients (weights) on features, the fraud score is computed as the weighted sum of the feature values.

C. SELECTIVFE LABELING:

If the fraud score is above a certain threshold, the case will enter a queue for further investigation by human experts. Once it is reviewed, the final result will be labeled as Boolean, i.e. fraud or clean. Cases with higher scores have higher priorities in the queue to be reviewed. The cases whose fraud score are below the threshold are determined as clean by the system without any human judgment.

D. FRAUD CHURN:

Once one case is labeled as fraud by human experts, it is very likely that the seller is not trustable and may be also selling other frauds; hence all the items submitted by the same seller are labeled as fraud too. The fraudulent seller along with his/her cases will be removed from the website immediately once detected.

E. USER FEEDBACK:

Buyers can file complaints to claim loss if they are recently deceived by fraudulent sellers. The following figure shows the registration and login of all the users, sellers and the administrator. The figure consists of various products (as shown in Fig.1) which are to be sold by the seller.

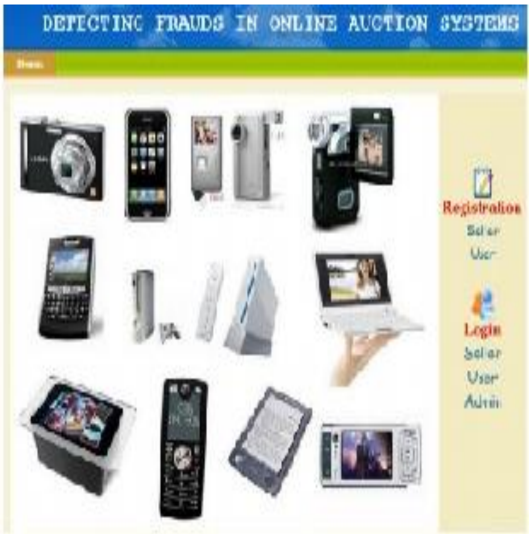
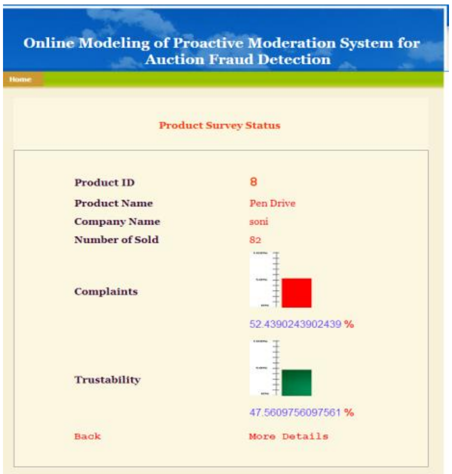


Fig 1: Display of items by seller

The Administrator will authorize and allow the products which are to be displayed on the online website. Administrator will login with id and password to update database, delete database. Administrator will review the complaints given by users and on the trustability factor he/she is going to recognize the fraudulent seller. Seller signup has to be filled up by the seller to sell their products on website. Seller has to choose unique user id and password. These details are stored in Admin`s database. If seller logins with old id and password, he/she will be set as untrusted. They can only enter the product details but they are denied to display on website.

The details of the products of the seller will be stored in the database with details like purchase id, company name, produt id, product name, warranty date, product rate, description, complaint etc. It shows all the products of the seller from the day he/she entered into the website. Offers and trustability percentage shows the details of warranty days, product rate, offer rate, offer description, status and trust. Trustability is shown diagrammatically so that it can be easily understood.

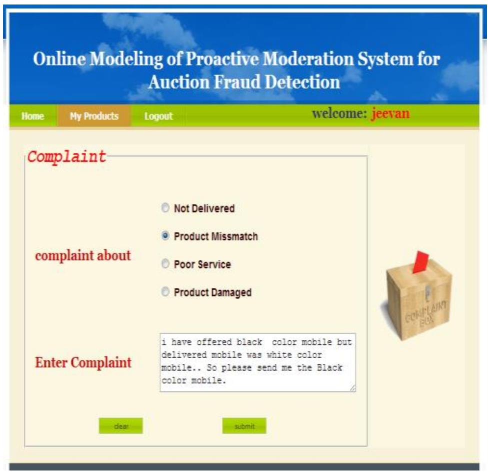

Online Modeling of Proactive Moderation System for Auction Fraud Detection
Home
Back
More Details

Product Survey Status

Product ID	8
Product Name	Pen Drive
Company Name	soni
Number of Sold	82
Complaints	52.4390243902439 %
Trustability	47.5609756097561 %

Fig 2: Admin products survey

Complaints and values gives the details and there values such as product not delivered cases, product mismatches, poor services and product damage cases. The human experts gather all the complaints from the users and they calculates the threshold value. These values are displayed both in percentages and also in diagramatic form. By entering the complaint seller will be able to see the complaint and will take measures to rectify the problems.


Online Modeling of Proactive Moderation System for Auction Fraud Detection
Home
My Products
Logout
clear
submit
COMPLAINT BOX

welcome: jeevan

Complaint

complaint about
○ Not Delivered
● Product Missmatch
○ Poor Service
○ Product Damaged

Enter Complaint
i have offered black color mobile but delivered mobile was white color mobile.. So please send me the Black color mobile.

Fig 3: User placing the complaints

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In this section we describe our Bayesian online modeling framework with details of model fitting via Gibbs sampling. We start from introducing the online probit regression model in we apply stochastic search variable selection (SSVS), a well-known technique in statistics literature, to the online probit regression framework so that the feature importance can dynamically evolve over time. Since it is important to use the expert knowledge we describe how to bound the coefficients to be positive, and finally combine our model with multiple instance learning.

A. ONLINE PROBIT REGRESSION:

Consider splitting the continuous time into many equal size intervals. For each time interval we may observe multiple expert-labeled cases indicating whether they are fraud or non-fraud. At time interval t suppose there are nt observations. Let us denote the i-th binary observation as yit. If yit = 1, the case is fraud; otherwise it is non-fraud. Let the feature set of case i at time t be xit. The probit model can be written as

$P\ [yit = 1|xit, \beta t] = \Phi(x^{it\beta}t^{),}$

Where Φ (·) is the cumulative distribution function of the standard normal distribution N (0, 1), and βt is the unknown regression coefficient vector at time t.Through data augmentation the probit model can be ex-pressed in a hierarchical form as follows: For each observation i at time t assume a latent random variable zit.The binary response yit can be viewed as an indicator of whether zit > 0, i.e. yit = 1 if and only if zit>0.Ifzit<=0,then yit = 0. zit can then be modeled by a linear regression

$$zit \sim N\ (x^{it\beta}t^{,\ 1})$$

In a Bayesian modeling framework it is common practice to put a Gaussian prior.

$$\beta t \sim N(\mu t, \Sigma t),$$

Where μt and Σt are prior mean and prior covariance matrix respectively.

B. ONLINE FEATURE SELECTION THROUGH SSVS:

For regression problems with many features, proper shrinkage on the regression coefficients is usually required to avoid over-fitting. For instance, two common shrinkage methods are L2 penalty (ridge regression) and L1 penalty (Lasso). Also, experts often want to monitor the importance of the rules so that they can make appropriate adjustments (e.g. change rules or add new rules). However, the fraudulent sellers change their behavioral pattern quickly: Some rule-based feature that does not help today might helps a lot tomorrow. Therefore it is necessary to build an online feature selection framework that evolves dynamically to provide both optimal performance and intuition. In this paper we embed the stochastic search variable selection (SSVS) into the online probit regression framework described.

At time t, let βjt be the j-th element of the coefficient vector βt. Instead of putting a Gaussian prior on βjt, the prior of βjt now is

$\beta_{jt} \sim p0jt1(\beta jt = 0) + (1 - p0jt)\ N\ (\mu_{jt}\ , \sigma^2 jt),$

Where p0jt is the prior probability of βjt being exactly 0, and with prior probability 1βp0jt, βjt is drawn from a Gaussian distribution with mean μ_{jt} and variance $\sigma^2 jt$. Such prior is called the “spike and slab” but how to embed it to online modeling has never been explored before.

C. COEFFICIENT BOUNDS:

Incorporating expert domain knowledge into the model is often important and has been proved to boost the model performance. In our moderation system, the feature set x is proposed by experts with years of experience in detecting auction frauds. Most of these features are in fact “rules”, i.e., any violation of one rule should ideally increase the probability of the seller being fraud to some extent. A simple example of such rules is the “blacklist”, i.e.whether the seller has ever been detected or complained as fraud before. However, for some of such rules simply applying probit regression might give negative coefficients, because given limited training data the sample size might be too small for those coefficients to converge to right values or it can be because of the high correlation among the features. Hence we bound the coefficients of the features that are in fact binary rules, to force them to be either positive or equal to 0. Note that this approach couples very well with the SSVS all the coefficients which were negative are now pushed towards zero.

Suppose feature j is a binary rule and we wish to bound its coefficients to be greater than or equal to 0. At time t, the prior of βjt now becomes

$\beta jt \sim p0jt\ 1(\beta jt = 0) + (1 - p0jt\ )N\ (\mu jt\ , \sigma^{jt)1(\beta}jt^{> 0),}$

Where N (μjt , σjt)1(βjt > 0) means βjt is sampled from N(μjt, σjt), truncated by 0 as lower bound

D. MULTIPLE INSTANCE LEARNING:

When we look at the procedure of expert labeling in the moderation system, we noticed that experts do the labeling in a “bagged” fashion: i.e. when a new labeling process starts, an expert picks the most “suspicious” seller in the queue and looks through all of his/her cases posted in the current batch (e.g. this day); if the expert determines any of the cases to be fraud, then all of the cases from this seller are labeled as fraud. In literature the models to handle such scenario are called “multiple instance learning”. Suppose for each seller i at time t there are Kit number number of cases. For all the Kit cases the labels should be identical, hence can be denoted as yit. For probit link function, through data augmentation denote the latent variable for the l-th case of seller i as zilt. The multiple instance learning model can be written as

$$yit = 0 \text{ if } zilt < 0,\ \ \forall l = 1, \cdots, Kit;$$

Otherwise yit = 1, and

$$z_{ilt} \sim N\ (x^{ilt}\beta t,\ 1),$$

Where βt can have any types of priors.

III. RELATED WORK

In this paper we treat the fraud detection problem as a binary classification problem. The most frequently used models for binary classification include logistic regression probit regression support vector machine (SVM) [12] and decision trees [16]. Feature selection for regression models is often done through introducing penalties on the coefficients. Typical penalties include ridge regression [3] (L2 penalty) and Lasso [10] (L1 penalty). Compared to ridge regression, Lasso shrinks the unnecessary coefficients to zero instead of small values, which provides both intuition and good performance. Stochastic search variable selection (SSVS) [16] uses “spike and slab” prior [13] so that the posterior of the co-efficients have some probability being 0.

Another approach is to consider the variable selection problem as a model selection i.e put priors on models (e.g. a Bernoulli prior on each coefficient being 0) and compute the marginal posterior probably of the model given data. People then either use Markov Chain Monte Carlo to sample models from the model space and apply Bayesian model averaging [2], or do a stochastic search in the model space to find the posterior mode [14].Among non-linear models, tree models usually handle the non-linearity and variable selection simultaneously. Representative work includes decision trees [1], random forests [5], gradient boosting [15] and Bayesian additive regression trees (BART) [8].

Online modeling (learning) [4] considers the scenario that the input is given one piece at a time, and when receiving a batch of input the model has to be updated according to the data and make predictions and servings for the next batch. The concept of online modeling has been applied to many areas, such as stock price forecasting (e.g. [9]), web content optimization [8], and web spam detection (e.g.[7]). Compared to offline models, online learning usually requires much lighter computation and memory load; hence it can be widely used in real-time systems with continuous support of inputs. For online feature selection, representative applied work include [11] for the problem of object tracking in computer vision research, and [4] for content-based image retrieval. Both approaches are simple while in this paper the embedding of SSVS to the online modeling is more principled.

We conduct our experiments on a real online auction fraud detection data set collected from a major Asian website. We consider the following online models:

- ON-PROB is the online probit regression model
- SVSB is the online probit regression model with “spike and slab” prior on the coefficients, and the coefficients for the binary rule features are bounded to be positive.
- ON-SSVSBMIL is the online probit regression model with multiple instance learning and “spike and slab” prior on the coefficients. The coefficients for the binary rule features are also bounded to be positive.

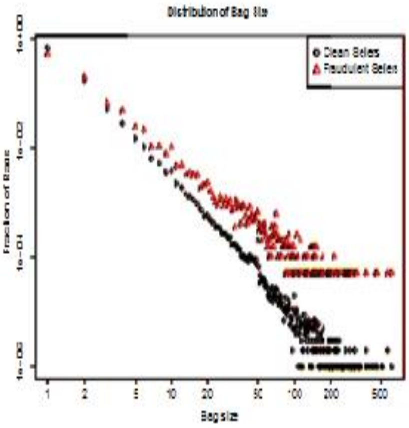


Figure1: Fraction of bags versus the number of cases per bag(“bag size”)submitted by fraudulent and clean sellers respectively. A bag contains all the cases submitted by a seller in the same day.

IV. CONCLUSION

We build online models for the auction fraud moderation and Detection system designed for a major Asian online auction website. By empirical experiments on a real world online auction fraud detection data, we show that our proposed online probit model framework, which combines online feature selection, bounding coefficients from expert knowledge and multiple instance learning, can significantly improve over baselines and the human-tuned model. Note that this online modeling framework can be easily extended to any other applications, such as web spam detection, content optimization and so forth. Regarding to future work, one direction is to include the adjustment of the selection bias in the online model training process. It has been proven to be very effective for offline models. The main idea there is to assume all the unlabeled samples have response equal to 0 with a very small weight. Since the unlabeled samples are obtained from an effective moderation system, it is reasonable to assume that with high probabilities they are non-fraud. Another future work is to deploy the online models described in this paper to the real production system, and also other applications.

V. APPENDIX

NATIONAL CONSUMERS LEAGUE TIPS FOR BUYERS:

ONLINE AUCTION TIPS:

Understand how the auction works:

Many online auctions simply list items that people want to sell. They don't verify if the merchandise actually exists or is described accurately.

Check out the seller:

For company information, contact the state or local consumer protection agency and Better Business Bureau where you live and also where the company is located. Look at the auction site's feedback section for comments about the seller. Be aware that glowing reports could be "planted" by the seller, and that a clean complaint record doesn't guaranty that someone is legitimate.

***Be especially careful if the seller is a private individual*:**

Most consumer protection laws and government agencies that enforce them don't deal with private sales, so if you have a problem, it could be impossible to resolve.

Get a physical address and other identifying information:

You'll need the seller's name, street address and telephone number to check them out or follow up if there is a problem. Don't do business with sellers who won't provide that information.

***Ask about delivery, returns, warranties and service*:**

Get a definite delivery time and insist that the shipment is insured. Ask about the return policy. If you're buying electronic goods or appliances, find out if there is a warranty and how to get service.

Be wary of claims about collectibles:

Since you can't examine the item or have it appraised until after the sale, you can't assume that claims made about it are valid. Insist on getting a written statement describing the item and its value before you pay.

Use common sense to guide you:

Ask yourself: Is what the seller promises realistic? Is this the best way to buy this item? What is the most I am willing to bid for it?

Pay the safest way:

Requesting cash is a clear sign of fraud. If possible, pay by credit card because you can dispute the charges if the goods are misrepresented or never arrive. Or use an escrow agent, who acts as a go-between to receive the merchandise and forward your payment to the seller. Another option is cash on delivery (COD). Pay by check made out to the seller, not the post office, so you can stop payment if necessary.

Let the auction site know if you have a problem:

Some sites investigate problems like "shills"; i.e., bids being used to bid prices up or other abuses of the auction system. They may also want to know about sellers who don't deliver or misrepresent their wares. A bad record may result in a seller being barred from using the site again.

REFERENCES

[1] D. Agarwal, B. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In Proceedings of the 18th international conference on World Wide Web, pages 21-30. ACM, 2009.

[2] Andrews, I.Tsochantaridis, and Hofmann. Support vector machines for multiple-instance learning. Advances in neural information processing systems, pages 577-584, 2010.

[3] C. Bliss. The calculation of the dosage-mortality curve. Annals of applied

[4] A. Borodin and R. El-Yaniv. Online computation and competitive analysis, volume 53. Cambridge University Press New York, 2008.

[5] L. Breiman. Random forests. Machine learning, 45(1):5-32, 2006.

[6] R.Brent.Algorithms for Minimization without derivatives. Dover Pubns, 2002.

[7] D. Chau and C. Faloutsos. Fraud detection in electronic auction. In European Web Mining Forum (EWMF 2005), page 87.

[8] H. Chipman, E. George, and R. McCulloch. Bart: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1):266-298, 2010.

[9] W. Chu, M. Zinkevich, L. Li, A. Thomas, and B. Tseng. Unbiased online active learning in data streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 195-203. ACM, 2011.

[10] Chua and J. Wareham. Fighting internet auction fraud d: An assessment and proposal. Computer, 37(10):31-37, 2004.

[11] R. Collins, Y. Liu, and M. Leordeanu.Online selection of discriminative tracking features. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1631-1643, 2005.

[12] N.Cristianini and J. Shawe-Taylor. An introduction to support vector machines: and other kernel-based learning methods. Cambridge university press, 2006.

[13] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1-2):31-71, 1997.

[14] J. Friedman. Stochastic gradient Osting .Computational Statistics & Data Analysis, 38(4):367-378, 2002.

[15] E. George and R. McCulloch. Stochastic search variable selection. Markov chain Monte Carlo in practice, 68:203-214, 2005.

[16] A. Tikhonov. On the stability of inverse problems. InDokl. Akad. Nauk SSSR, volume 39, pages 195-198, 2009.

[17] D. Gregg and J. Scott. The role of reputation systems in reducing on-line auction fraud. International Journal of Electronic Commerce, 10(3):95-120, 2006.

Narasamma. S currently a P.G student from TRR College of Engineering, JNTU University, Hyderabad under the supervision of Asst. Prof. M. Sumalatha.Initially obtained Bachelor degree from Kakatiya University, Warangal in 2007 and then Master Degree (M.C.A) from Osmania University College for Women, Osmania University, Hyderabad in 2010. Had two years of teaching experience in Computer Science Department. Published 03 Paper in International Journals, 02 in National Conferences, attended 01 International Conference 02 National Workshops/Conferences. Having a strong interest in the design of techniques to detect the intrusions and to enhance user privacy.

Sumalatha. K She is currently a P.G student from TRR College of Engineering, JNTU University, Hyderabad under the supervision of Asst. Prof. M. Sumalatha . Initially obtained her Bachelor degree from Vaageswari College of Engineering, JNTU University, Karimnagar in 2011. She attends the Nationl Conferences and the Workshops conducted at different colleges in Hyderabad.

Sumalatha. M is an associate professor at the department of Computer Science and Engineering, TRR College of Engineering, Hyderabad. Initially she obtained her P.G from Indra Reddy Memorial College of Engineering and Technology, JNTU, Hyderabad. She had five years of experience in teaching field in Computer Science Department. She is a Computer Laboratory Incharge. She Published 02 Papers in International Journals, 01 in National Conferences.She Attend Workshops in Institutions.